4

Proteomics as a Tool for the Characterization of Microbial Isolates and Complex Communities

Florence Arsène-Ploetze[1], Christine Carapito[2], Frédéric Plewniak[1] and Philippe N. Bertin[1*]
[1]*Génétique moléculaire, Génomique et Microbiologie, Université de Strasbourg, Strasbourg*
[2]*Laboratoire de Spectrométrie de Masse Bio-Organique, Institut Pluridisciplinaire Hubert Curien, Strasbourg*
France

1. Introduction

Proteins may be considered as the main effectors of biological responses of organisms to specific environmental conditions, instead of messenger RNAs. Indeed, a modulation in their activity does not always depend on a modified expression of the corresponding genes but rather on post-translational modifications. Proteome analysis may therefore constitute an appropriate approach to address the question of organism adaptation to environmental stresses or growth under extreme conditions. Recently, the knowledge of the organisms' physiology has led to deep changes in the investigation methods, favouring the use of global analysis methods in complement with conventional strategies. Instead of studying individual proteins or metabolic products, the integral profile of organisms can now be established. This may be of importance when studying adaptive and stress responses in microorganisms because of their multifactorial character. In particular, the differential analysis in various growth conditions of the whole protein content (« proteome »), which allows the simultaneous quantification of gene products in an organism, represents part of the so-called integrative biology (Bertin et al., 2008).
Genomics is a conceptual approach that aims to study the biology of microorganisms by analysing the complete genetic information they contain. This scientific discipline really emerged more than fifteen years ago with the characterization of the first complete genome of autonomous organisms (Bertin et al., 2008). An important reduction of sequencing costs associated with new high-throughput technologies has led to an explosion of genomic programs that now concern organisms in all domains of life (http://www.genomeonline.org). Most of the descriptive and functional genomic efforts initially focused on human pathogens, such as bacteria and parasites, and next, on higher eukaryotes. More recently, there has been a growing interest in microorganisms isolated from various habitats, including extreme ecological niches, to characterize specific properties that allow these organisms to grow in such environments. The field of application of proteomics thus expended in line with genomics. These works should lead not only to a

* Corresponding Author

better understanding of ecosystems themselves, but also to the identification of novel functions that may be exploitable for biotechnological applications, in particular in the bioremediation of contaminated environments. A better understanding of the involved elements, their spatial and temporal distribution, the metabolic pathways they belong to, would allow drawing an integrated picture of biological processes under study. This could lead to an optimal use of microorganism properties, favouring the desired effects. In this review, global proteomics approaches allowing deciphering the physiology of one microorganism or the functioning of a community will be presented, as well as recent advances in targeted proteomics approaches. Finally, the huge amount of data generated by such approaches needs integrative analyses that require specific proteome databases.

2. Global proteomics approaches

In their natural habitat, microorganisms rapidly adapt to environmental changes by modulating their protein content or activity, for instance via post-translational modifications. Therefore, to highlight the physiological state of a microorganism in one particular condition, a large-scale study of its proteome is a widespread approach. In such a workflow, the establishment of global protein profiles (Figure 1) requires protein extraction, separation steps that are often obtained by two-dimensional gel electrophoresis (2DE) followed by mass spectrometry analysis for protein identification.

2.1 Proteomics methodology: From sample preparation to protein identification

Protein extraction and separation in a homogenous population require first to optimize the lysis conditions. Sample preparation is a fundamental step and several protocols are usually tested, such as those described in (Cañas et al., 2007). Physical lysis methods are the most useful methods in the case of microorganisms: vortexing and grinding with glass beads, sonication, freeze/thaw or alternating cycles of high and low pressure. Combining these mechanical methods with enzymatic lysis or use of detergents may improve cell lysis efficiency. 2DE separation has shown to be one of the most common separation techniques used in proteomic studies (Rabilloud et al., 2010). Proteins are separated in a first step according to their charge and in a second step according to their molecular weight. They are then usually visualised by an organic dye (Coomassie blue), by metallic salt reduction (silver nitrate) or fluorescent labelling (Sypro, DeepPurple…). Bidimensional proteome analysis presents however several limitations. Indeed, whatever the detection method used, all proteins of any organism cannot be visualised because some of them are present at very low levels. In addition, some proteins are quite unstable while others are less labile. Moreover, membrane proteins are usually more difficult to detect on 2D gels because of their low solubility. Therefore, other separation techniques may be used such as monodimensional SDS-PAGE, in particular to retain the membrane proteins (Laemmli, 1970), non-gel strategies such as MudPIT approaches (multidimensional protein identification technology) (Fränzel & Wolters, 2011) or any other liquid or affinity chromatography-based separations (Gundry et al., 2009). Many kinds of original chromatography types and combinations have been explored in the field of proteomics to fractionate and separate complex protein mixtures prior to mass spectrometry (MS) analysis either at a protein or at a peptide level. Each of those approaches presents advantages and drawbacks and the choice of the separation method used is a crucial step in the proteomics workflow. Overall, the higher success of one or the other separation method is highly sample-dependent.

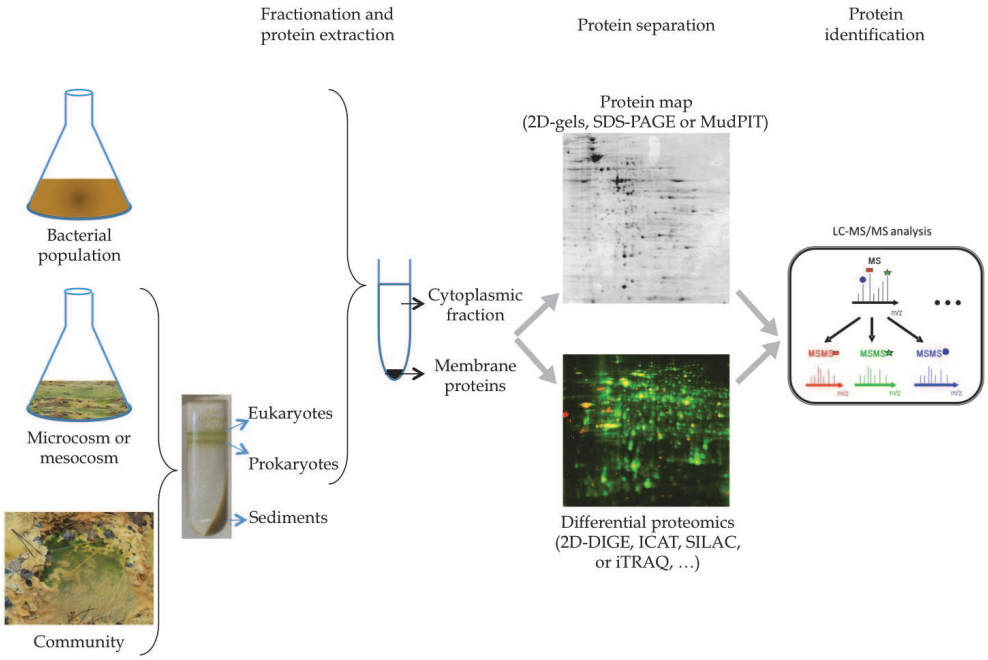


Fig. 1. Classical proteomics workflow to study the physiology of microbial isolates or complex communities.

Once separated, proteins are identified by mass spectrometry. The recent development of functional genomics approaches has led to considerable progress in identification methods (Casado-Vela et al., 2011). Proteins are characterized by mass spectrometers able to ionize and precisely determine the masses of ionized molecules. Proteins of interest are recovered from gels or from any other chromatography separation and enzymatically digested, e.g. by trypsin which specifically cuts the polypeptidic chain at lysine or arginine residues. The whole set of generated peptides are then analysed by MS and most commonly tandem MS (MS/MS) to precisely measure the molecular weight of the peptides and their associated fragments (in MS/MS mode). The experimental MS data are compared to theoretical data calculated from the available protein sequence databases derived from the genome sequence. Historically, the Peptide Mass Fingerprint (PMF) approach, based on the measurement of the peptide masses only, was used to identify proteins (mostly by Matrix-Assisted Laser Desorption Ionization Time of Flight MS, MALDI-TOF-MS) but this approach quickly revealed to be insufficiently specific with the exponentially growing protein sequence databases. MS/MS approaches nowadays constitute the standard method to reliably identify proteins. Over the last 10 years, numerous algorithms, proprietary or open-source, have been developed to compare and score the matching between experimental MS/MS data and theoretical mass lists calculated from expected protein sequences (several of the most commonly used tools are listed in Table 1) (Nesvizhskii, 2010). The MS/MS protein identification workflow is now well established and allows the performance of high throughput and large scale proteomic experiments provided that the genome of the studied organism is sequenced.

Database	Access
Mascot	http://www.matrixscience.com
Phenyx	http://www.genebio.com/products/phenyx
OMSSA	http://pubchem.ncbi.nlm.nih.gov/omssa
X! Tandem	http://www.thegpm.org/TANDEM/

Table 1. Tools useful for MS/MS data analysis.

However, even when genomic information is available, protein identification may be complicated by lacks/errors in the predicted protein sequence databases introduced by automatic genome annotation (translational frameshift, read-through of stop codons) or by post-translational modifications (e.g., glycosylation or phosphorylation) hindering the mass matching procedure. To circumvent those errors widespread in non reference and not thoroughly annotated genomes, original alternative identification strategies have been developed which use the complete unannotated genome sequence to interpret the MS/MS data. These approaches have opened the avenue to proteogenomics, defined as the use of proteomics results to enhance the knowledge of the genome (Delalande et al., 2005; Gallien et al., 2009). Finally, when the genome of the organism under study has not yet been sequenced, *de novo* sequencing is mandatory. This consists in interpreting individually each high quality MS/MS spectrum to derive amino acid sequence tags. These sequence tags are then submitted to MS-BLAST (http://dove.embl-heidelberg.de/Blast2/msblast.html) homology searches in order to identify the proteins by sequence homology with orthologous proteins present in the databases (Carapito et al., 2006).
Altogether, the recent developments in proteomics, in particular in MS instrumentation to gain sensitivity, resolution and mass accuracy, as well as the important increase of genomic data enabled proteomics to become a widespread, useful and robust technique to understand the adaptive capacities of microorganisms, under laboratory conditions but also in their natural habitat within complex communities.

2.2 Proteomics as a tool to understand the physiology of environmental isolates

Proteomics has two major goals. On the one hand, proteomic maps can first make an inventory of functions expressed in an organism under specific conditions. On the other hand, differential proteomic analyses allow studying the response of microorganisms to changes in the environment as well as the underlying regulatory mechanisms. Several examples of these two strategies allowing better understanding of the physiology of environmental isolates are presented in the following sections.
First, by using 2DE or SDS-PAGE separation techniques, the global or partial proteome maps (cytoplasmic, membrane or extracellular fraction) of several microorganisms have been drawn. These often concern model organisms, e.g. *B. subtilis* (Hecker & Völker, 2004) or human pathogens, e.g. *Mycobacterium tuberculosis*, the etiologic agent of tuberculosis (Schmidt et al., 2004). This approach was recently used to list proteins expressed by an arsenic resistant bacterium, *Herminiimonas arsenicoxydans*, which is able to resist and grow in harsh conditions, particularly in the presence of arsenite (Weiss et al., 2009). Another example of bacterium able to adapt to extreme conditions is *Deinococcus geothermalis,* found

in geothermal wells. In this bacterium, cytosolic and cell envelope proteome maps revealed that one-fourth of the cell envelope proteome corresponds to *Deinococcus* specific proteins such as V-type ATPases, and that repair enzymes are highly expressed and among the most abundant proteins, even in the absence of stress (Liedert et al., 2010). Finally, these techniques may be used to focus on a particular fraction of the proteome. As an example, using specific staining procedures, a 2DE map of iron-metalloproteins has been drawn in the acidophilic archaeon *Ferroplasma acidiphilum* (Ferrer et al., 2007). Remarkably, the results suggest that the high content of metalloproteins present in this organism represents a relic of early life on Earth, where metals were abundant because of widespread volcanic and hydrothermal activities.

Second, to get further insight into the adaptation capacities of microorganisms, differential proteomic analyses characterize proteins which expression is induced or repressed in response to a particular stimulus, and defines thus a stimulon. Using 2DE, the amount of proteins expressed in different conditions or backgrounds may be compared, which requires robust quantification of each protein in each condition. The use of 2DE coupled to fluorescent labelling (DIGE) makes such quantification easier (Yan et al., 2002). This differential proteomic strategy was extensively used to decipher the adaptive response of pathogens or model bacteria such as *Bacillus subtilis* (Bertin et al., 2008; Hecker & Völker, 2004; Jungblut, 2001). Recently, the physiology of an increasing number of environmental isolates has been studied by proteomics approaches. For example, the adaptation to cold has been studied in psychrophilic microorganisms, such as *Pseudoalteromonas haloplanktis* (Piette et al., 2010) and in the archeon *Methanococcoides burtonii* (Saunders et al., 2005). Similarly, arsenic bacterial metabolism has been investigated in *H. arsenicoxydans* (Carapito et al., 2006; Muller et al., 2007) and *Thiomonas* sp. (Bryan et al., 2009). In *H. arsenicoxydans*, in addition to the proteome map listing the proteins expressed in the presence of arsenite (see above), differential proteomic analyses data were combined with transcriptomics data to study its adaptive response in the presence of arsenite (Cleiss-Arnold et al., 2010; Muller et al., 2007; Weiss et al., 2009). These methodologies revealed that this bacterium is able to grow in the presence of arsenic by inducing the expression of proteins involved in several processes such as oxidative stress, arsenic resistance, energy metabolism or motility/chemotactism. Differential proteomics experiments performed on *Thiomonas* allowed comparing the arsenic response in several strains. Indeed, proteomics has highlighted metabolic differences between *Thiomonas arsenitoxydans* 3As and *Thiomonas arsenivorans* strains. In the presence of As(III), proteins involved in carbon fixation were shown to be preferentially accumulated in *Tm. arsenivorans* but less abundant in *Tm. arsenitoxydans 3As*, supporting the hypothesis that *Tm. arsenivorans* is capable of optimal autotrophic growth in the presence of As(III) when used as an inorganic electron donor. One response shared by these arsenic-oxidizing bacteria is the induction in the presence of arsenite of phosphate transporters, as well as proteins involved in glutathione metabolism, DNA repair and protection against oxidative stress (Bryan et al., 2009; Carapito et al., 2006; Cleiss-Arnold et al., 2010; Weiss et al., 2009).

In differential analyses, the 2DE technology has however one particular limitation, i.e. several proteins may be resolved in the same spot hindering their respective quantification. To avoid such a problem, differential protein patterns can be identified using non-gel protein separations coupled with isotope labelling approaches. To find proteins or peptides with significant differences of concentrations in sampled proteomes, different stable heavy isotope labelling techniques can be applied like ICAT, SILAC,

iTRAQ or ICPL. Depending on the sample origin, isotope labelling may be performed on different levels (organism, cell, protein, or peptide) and on different reactive groups (Gevaert et al., 2008). As an example, 2DE and ICAT approaches were combined to study the aromatic catabolic pathways in *Pseudomonas putida* KT 2440. Interestingly, it appears that these two methods are complementary since 110 and 80 proteins were shown to be induced in the presence of benzoate, using ICAT or 2DE, respectively, and only 19 common proteins were identified using both approaches (Kim et al., 2006). Even though those approaches have proven to allow precise quantification of numerous proteins in various applications, each of them presents drawbacks and limitations. For instance, stable isotope labelling with amino acids in cell culture (SILAC) is limited to applications dealing with proteins obtained from cell cultures, free amino-group labelling (like ICPL) induces significant increase of sample complexity leading to an aggravation of undersampling problems, isobaric labelling (like iTRAQ) requires high resolution MS/MS data to be acquired and is often unsuitable for most widespread ion trap MS/MS. The choice of the approach to apply for quantification is therefore very sample-dependent and crucial for the success of the proteomics application.

Once adaptation capacities have been identified, it can be interesting to understand the regulation network allowing microorganisms to respond quickly to changes in their environment. With such an aim, differential proteomics is useful to decipher the role of global regulators and to list genes belonging to the same regulon, i.e. genes that are regulated by the same regulator. Such approaches have helped to highlight unsuspected regulatory networks, revealing that some bacteria have developed sophisticated mechanisms to survive or grow in a large panel of conditions. As an example, the global Crc protein that control the metabolism of carbon sources and catabolite repression of *Pseudomonas aeruginosa* was shown to be involved in the regulation of virulence gene expression (Linares et al., 2010). Finally, proteomics can also be used to highlight post-translational modifications (PTMs) that may be crucial for rapid regulation of protein activity. For instance, a phosphoproteomic study allowed identifying kinases involved in the regulation of several cellular processes in bacteria (Grangeasse et al., 2010).

2.3 Proteomics as a tool to understand the functioning of communities: Metaproteomics or environmental proteomics

Studying microorganisms in laboratory conditions may not reflect their particular adaptation capacities in their environmental niches. For example, in the case of pathogens, symbionts or commensals, it is crucial to identify not only proteins expressed in response to abiotic changes, but also in response to biotic factors, such as those expressed by their host. The major difficulty in such studies is to distinguish microbial and host proteins, a difficulty that is reduced when the genome of both organisms is known. The second problem is to extract sufficient amount of microbial proteins in order to detect them. Recent advances in protein identification have allowed access to such information (see below), as shown by the identification of key proteins involved in virulence in several pathogens such as *Echinococcus granulosus* metacestode (Monteiro et al., 2010), *Clostridium perfringens* (Sengupta & Alam, 2011) or *Anaplasma* when present in the tick vector (Ramabu et al., 2010). Similarly, proteomics has been used to address the complex processes governing the interactions between symbiotic microorganisms and their host and *vice versa*, e.g. the adaptive response of plants interacting with mycorrhizae (Bona et

al., 2011). Recently, proteomics approaches have been developed to study the interactions of microorganisms with their host or microbial communities that may contain uncultured microbes, thus extending our knowledge of the diversity of microbial metabolic processes. Microbial communities are complex biological assemblies whose study has been difficult for a long time because of the inability to culture many of their components. However, the taxonomic diversity studies performed by the analysis of 16S rRNA sequences suggest that in any given environment only a small fraction of organisms present can actually be cultivated. These communities can now be explored as a whole by the sequencing of their genomic DNA content, i.e. their metagenome (Bertin et al., 2008). In parallel, a novel proteomic approach called metaproteomics or environmental proteomics has emerged to characterize in a global way the protein content of microbial communities. The metaproteomics approach has some advantages, compared to other functional genomic approaches, such as metatranscriptomics, i.e. the study of mRNA expressed by a community. Indeed, as proteins are more stable than RNA (especially those originating from prokaryotes), the metaproteome content is supposed to be less affected by the extraction procedure, and probably gives a better insight into the biological functions expressed *in situ*. Several examples of recent studies in this environmental proteomics field are presented below.

2.3.1 Environmental proteomics as a tool to characterize uncultured microorganisms: Advantages and limitations

Interestingly, the metaproteomics approach may give taxonomic information complementary to the 16S rRNA gene-based approach, commonly used to analyse the community structure. Previous observations established that it is not always possible to affiliate some bacteria using only 16S rRNA gene sequences (Schleifer, 2009). Indeed, some microorganisms showing very similar 16S rRNA gene sequences turned out to belong to different taxa when other phylogenetic markers were used. For instance, in a recent study, metaproteomics highlighted the expression of proteins involved in conserved biological processes that could be assigned to a specific taxon, at the genus or species level, whereas the RDP (Ribosomal Database Project) analysis allowed the affiliation of only 28% at the genus level (Halter et al., 2011). More generally, the identification of signature peptides in orthologs has enabled the use some proteins as taxonomic markers to describe the active community in the Carnoulès AMD (Bruneel et al., 2011), and to differentiate various ecotypes present in a similar ecosystem (Simmons et al., 2008).

In addition, metaproteomics has enabled the analysis of the role of uncultured microorganisms *in situ*. For example, a study of microbes growing in water plant sludge led to the identification of several proteins belonging to an uncultured organism of the *Rhodocyclus* lineage known to accumulate polyphosphates (Wilmes & Bond, 2004). Similarly, proteins synthesized by microorganisms flourishing in an AMD ecosystem, i.e. inside a biofilm (Ram et al., 2005), have been inventoried. In this study and others, strain-resolved expression patterns indicated that microorganisms belonging to the same species with less than 1% divergence in nucleotide sequences of genes encoding 16S rRNA (ecotypes) coexist in ecosystems. At a functional level, this microdiversity can lead to functional diversity, since these strains may play distinct roles (Denef et al., 2010a, 2010b). In another study, synergistic/antagonistic interactions between fungi and Rhizobacteria were explored (Moretti et al., 2010). Proteomic patterns of a microbial consortium were

compared in the presence or the absence of antibiotic, in order to evaluate the bacterial impact on the consortium functioning. Using this strategy, candidate proteins were identified that may provide advantages for the consortium to out-compete pathogen strains such as *Fusarium*.

In some cases, the high level of diversity makes the metaproteomics approach rather difficult to apply. A low number of identified proteins have been observed previously in soil or sediments where a high level of diversity was observed (Benndorf et al., 2007; Halter et al., 2011; Taylor & Williams, 2010). As pointed out by the authors of recent reviews, metaproteomics studies are successful when applied to communities with low levels of diversity. When a high level of diversity is observed, each protein is diluted in a complex mixture and only the most abundant proteins are therefore likely to be identified. Moreover, a large proportion of the bacteria forming such a community have usually never been studied so far *in vitro* and their genome sequences, and hence their protein sequences, which are required for MS identification, are not available in public databases. For example, unlike *Proteobacteria*, only a few *Acidobacteria* or archaeal protein sequences are available in the existing protein sequence databases. To prevent such a limitation, metaproteomics and metagenomics are nowadays often combined.

When the community is too complex or when this community is found in solid phase such as soil or sediments, it may be necessary to fractionate cells, in order to study only a fraction of the community (Figure 1). For instance, metaproteomes of key microbial populations, i.e. *Synechococcus* cells, were drawn after cells separation using microwave fixation and flow cytometric sorting (Mary et al., 2010). In another study, using density gradient, it was possible to separate microorganisms from sediments but also bacteria from the eukaryotic population and to study both populations separately. Indeed, in the Carnoulès arsenic-rich ecosystem, the bacterial community analyses revealed that proteins involved in the biomineralization of iron and arsenic were shown to be expressed by *Acidothiobacillus ferrooxidans* and *Thiomonas,* respectively, which supports a major role of these microorganisms in the natural attenuation of this highly contaminated environment (Bertin et al., 2011). This approach also revealed that most proteins were expressed by uncultured microorganisms belonging to a novel phylum, i.e. "*Candidatus* Fodinabacter communificans". These bacteria may play an indirect but important role in the functioning of the ecosystem by recycling organic matter or providing other members with cofactors such as vitamins (Bertin et al., 2011). An additional study revealed that *Euglena mutabilis*, an abundant protist found in this AMD as well as in other AMDs, produces organic compounds that could serve as nutrients for bacteria (Halter et al., unpublished).

2.3.2 Environmental proteomics as a tool to understand the dynamic and the functioning of ecosystems

To study factors that may influence the community adaptation, metaproteomics approaches are sometimes performed on controlled microcosms (Figure 1). For example, such an approach has been successfully used to study the temporal dynamics of microbial communities subjected to cadmium exposure and to characterize the resulting response in terms of toxicity and resistance (Lacerda et al., 2007). This study illustrates that metaproteomics can be used, not only to describe an ecosystem, but also to study its response to perturbations. For example, the spatial dynamics of bacterioplankton was evaluated along the Chesapeake Bay, the largest estuary in the United States, and the

proteins identified were shown to correlate with major microbial lineages, i.e. Bacteroides and Alphaproteobacteria, present in this ecosystem (Kan et al., 2005). Environmental proteomics combined with physiology and geochemical data allowed a description of the ecological distribution of dominant and less abundant organisms, and the changes along environmental gradients in a biofilm within the Richmond mine at Iron Mountain, California, or the effect of the pH on acidophilic AMD microbial communities (Belnap et al., 2011; Mueller et al., 2010). Other studies aimed to elucidate the *Geobacter* physiology during stimulated uranium bioremediation (Callister et al., 2010; Wilkins et al., 2009), or the community responses to different nutrient concentrations on an oceanic scale (Morris et al., 2010). In this study, a shift in nutrient utilization and energy transduction along a natural nutrient concentration gradient was observed, with a dominance of TonB-dependent transporters expressed in these samples. Although it is likely that only the dominant organisms will be visible in metaproteomic studies, results provide evidence that such an approach presents a considerable interest towards a comprehensive analysis of microbial ecosystems. In the future, such approaches will be improved in order to access a large amount of proteins expressed by individual cells within a community.

3. The potential of targeted proteomics approaches

The last 10 years, global proteomic approaches have allowed drawing long lists of hundreds to thousands of identified proteins in all types of proteome fractions. The major weakness of those lists is often the lack of quantitative data for the identified proteins, especially in the case of metaproteomic studies where quantification may be difficult. To alleviate this problem, the proteomics specialists recently initiated a paradigm shift from global approaches towards targeted approaches, trying to find ways to get better quantitative data even if one has to focus on a limited number of proteins of interest. Selected Reaction Monitoring (SRM)-based strategies appear to be the most promising approaches to reach this goal (Picotti et al., 2010) and applications, mostly in the field of protein biomarker research, are starting to become successful (Elschenbroich & Kislinger, 2011).

3.1 Selected Reaction Monitoring-based proteomics workflow

The general SRM-based workflow is described in Figure 2. The first step of a targeted proteomics experiment resides in the definition of a restricted list of proteins of interest. This is the major difference between global approaches (in which the goal is to identify the highest number of peptides/proteins) and targeted approaches in which the targets to focus on have to be defined prior to the experiment itself. Once the targets are defined, developing a SRM-based quantification method relies on the choice of a small series of peptides that will be used as tracers for each protein to be quantified. In the SRM scanning mode, the precursor ion corresponding to the targeted peptide is selected in the first mass filter (Q1) before entering the collision cell (q2) where it undergoes collision-induced dissociation. One fragment ion is then selected in the second mass filter (Q3) and its intensity is monitored. An ion pair of precursor/fragment ions is called a transition and several transitions are recorded for each targeted peptide. The critical steps of the method setup reside in the choice of those so-called proteotypic peptides (unique for the protein and visible in MS) to be used for each protein, in the selection of transitions (number and fragment types) to be followed for each of the selected peptides and, finally, in the

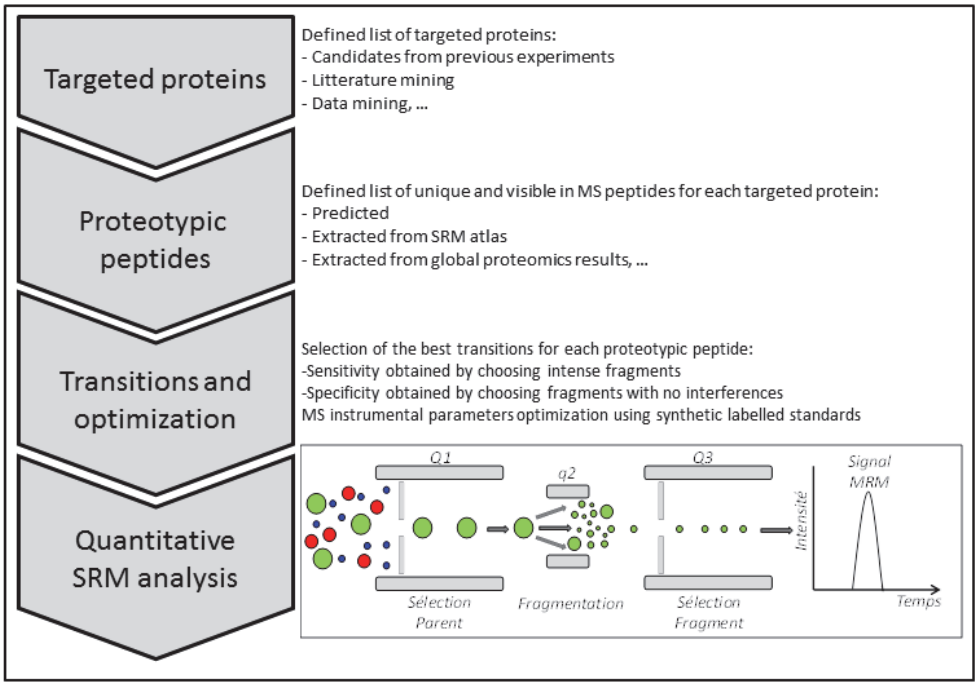


Fig. 2. Targeted SRM-based proteomics workflow

optimisation of the MS instrument parameters. The transitions have to be selected to offer both best sensitivity (intense fragments) and best selectivity (no interferences with other fragments). To accelerate these limiting steps in terms of time and cost, public libraries (atlases) of transitions are being constructed using synthetic peptides for a few proteomes of reference organisms, of which yeast and human (http://www.srmatlas.org). Those atlases will significantly facilitate the choice of the proteotypic peptides as they will contain the lists of 5 peptides for each predicted protein of the reference proteome, along with their optimal transitions and information on instrument parameters. Once the peptides and transitions are established, isotopically heavy labeled standards are required and need to be spiked into the samples in order to be able to quantify the endogenous peptides of interest by calculating heavy/light ratios (Gallien et al., 2009; Lange et al., 2008). The hypothesis-driven nature of such experiments overcomes the bias towards most abundant components and has already allowed previously unreached sensitivity levels using MS techniques.

3.2 SRM quantification in proteomics

The quantitative power of SRM mass spectrometry no longer needs to be proven. This approach has been used for a number of years for the quantification of small molecules such as metabolites of xenobiotics, hormones or pesticides with great precision (CV<5%). However, three main hurdles have hindered its application for the quantification of peptides and proteins. The first major hurdle to be overcome was sensitivity, or more

precisely, the capacity to quantify proteins of very low abundance in mixtures in which protein concentrations range over 5 to 10 orders of magnitude, even up to 12 orders in the case of plasma samples. To circumvent this problem, the introduction of fractionation methods permits considerable reduction of the quantification limit provided they are perfectly controlled and reproducible. For instance, SRM methods recently allowed the detection of concentrations typical for candidate protein biomarkers whose abundance can be as low as a few ng/ml or even hundreds of pg/ml in human plasma (Keshishian et al., 2009). The second hurdle was the reproducibility of SRM analyses for proteomics. This hurdle appears to have been overcome today: indeed, as part of a study carried out by the Clinical Proteomic Technology Assessment For Cancer Network Project (CPTAC, (Addona et al., 2009), interlaboratory CVs (including both variations due to sample preparations and MS analysis) between 10 and 23% were obtained across 9 different laboratories. Finally, multiplexing was made possible thanks to significant progress in electronics and acquisition and data processing software developed on triple quadrupole-type instruments. Today the simultaneous quantification in a single analysis of about a hundred peptides can be envisaged using a few hundred transitions.

3.3 SRM-strategies for quantifying proteins in microbial isolates and complex communities

So far, most of the SRM-based applications have dealt with biomarker studies and clinical proteomics (Gallien et al., 2009; Hüttenhain et al., 2009). Nevertheless, it has been proven to be very successfully applicable on *S. cerevisiae* and other whole proteome digests (Picotti et al., 2009). Even though microbial communities are extremely complex protein mixtures, both in terms of number of proteins and dynamic range, protein concentrations ranging over 12 orders of magnitude in the case of plasma samples have revealed the success of the technology. It is therefore reasonable to predict a widespread application of SRM quantification methods in many fields. Actually, a recent study has already demonstrated the possibility to absolutely quantify proteins in complex environmental samples and mixed microbial communities (Werner et al., 2009). An absolute prerequisite for the success of the method is a precise control and reproducibility of the sample preparation and fractionation steps. Additionally, one of the key factors for a reliable quantification will be the use of appropriate quantification standards. Indeed, the use of isotopically labelled standards considerably improves quantification reliability. In any case, absolute quantification of peptides, and thus the proteins producing them, is only possible through the simultaneous LC-SRM measurement of endogenous peptides and isotopically labelled standards added in known quantities. Several choices are possible: synthetic peptides (the AQUA method, Gerber et al. 2003), concatemers of peptides (the QconCAT method, (Beynon et al., 2005)) and protein standards, biochemically identical to the natural proteins to be assayed (the PSAQ method, (Brun et al., 2007)).

4. Data integration and proteome databases

High-throughput proteomics is a rapidly developing field enabling analyses at system level from complexes or cells and organs to environmental communities (Cannon & Webb-Robertson, 2007). With the huge amount of data produced by experiments and subsequent analyses, proteomics repositories will have to take up the challenge of large-scale storage, fast access and easy data retrieval. Furthermore, with the development of Systems Biology,

proteomics databases, like other "omics" databases, will require exchange and communication standards which could help to integrate data with related information from other databases or fields (genomics, genetics, metabolomics) into a wider scope. Several proteomics repositories have been established thus far and range from large-scale general databases to more specialized ones (Table 2). However, although some repositories exist that are specialized in microbiology-related proteomics, there is still a lack of proteomics databases dedicated to environmental microbiology.

Database	Description	Access
Swiss 2D-PAGE	1DE and 2DE data	http://world-2dpage.expasy.org/swiss-2dpage/
World 2D-PAGE portal	Federation of 2DE-based databases	http://world-2dpage.expasy.org/portal/
InPACT	Gel-based environmental microbiology database	http://inpact.u-strasbg.fr/
Proteome Database for Microbial Research	Gel and mass spectrometry microbiology database	http://www.mpiib-berlin.mpg.de/2D-PAGE/
GPMDB	Comprehensive mass spectrometry database for validation of identification and protein coverage	http://gpmdb.thegpm.org/
PeptitdeAtlas	Compendium of raw data coming from high-throughput proteomics technologies aiming at the annotation of eukaryotic genomes through a thorough validation of expressed proteins	http://www.peptideatlas.org/
PRIDE	Comprehensive mass spectrometry-derived peptide and protein identifications, MS mass spectra, and associated metadata.	http://www.ebi.ac.uk/pride/
Proteome Commons	Communautary resource for collaborative research and sharing of proteomics data	https://proteomecommons.org/

Table 2. Proteomics repositories useful for the analysis of microbial proteomes.

4.1 Technical challenges

4.1.1 Data storage and the scalability challenge

High sensitivity and high quality mass spectra are delivered by mass spectrometers at an ever-faster rate, yielding an ever-increasing amount of data. For instance, the data available from Proteome Commons (https://proteomecommons.org/) repository has currently (July 2011) a size of 16.7 TB comprising 12,895,832 data files (for the sake of comparison, the 1638

uncompressed flat files of Genbank release 184.0 require approximately 540 GB). Therefore, scalability is expected to become a major issue for those comprehensive proteomics repositories and high capacity storage and efficient data access may require the use of distributed IT technologies similar to those serving large databases on the web (Facebook Cassandra, Google BigTable, Amazon, Dynamo). As a matter of fact, in order to handle data and to facilitate access by users, Proteome Commons based its repository on an implementation of Tranche (https://trancheproject.org/), a free open-source file distributed storage and dissemination software (Falkner et al., 2008).

4.1.2 Data access: Needs for standards

One of the first efforts towards a single access point to proteomics data was the publication in 1996 of guidelines for building federated 2DE databases (Appel et al., 1996). Since then, ExPASy has developed Make2D-DB II, an environment to create, convert, publish, interconnect and keep up-to-date 2DE databases (Mostaguir et al., 2003). More recently, the ProteomeExchange consortium has been established to provide a single point of submission to PRIDE (Vizcaíno et al., 2009), PeptideAtlas (Deutsch et al., 2008) and Tranche (https://proteomecommons.org/tranche/) repositories. This consortium encourages the data exchange and sharing of identifiers between repositories so that the community may easily find datasets. Furthermore, in order to be effective, computational analysis and data-mining require the use of controlled vocabularies or ontologies for data descriptions. The Protein Ontology (Natale et al., 2010) provides a standardized vocabulary for the description of protein evolutionary relatedness (ProEvo), protein forms including isoforms or PTMs (ProForm) and protein-containing complexes (ProComp). On a larger scale, the HUPO proteomics standards initiative (HUPO-PSI) is aiming to define standards to ease data exchange and minimize data loss (Orchard & Hermjakob, 2007) in key areas of proteomics: protein separation, gel electrophoresis, mass spectrometry, molecular interactions, protein modifications and proteomics informatics. The HUPO-PSI is developing the minimum information about proteomics experiments (MIAPE) guidelines defining which information should minimally be reported about a proteomics experiment to allow critical assessment. It also develops data formats for capturing, describing and exchanging MIAPE-compliant data as well as supporting controlled vocabularies. Some of these standards, like MIAPE (Taylor et al., 2007), FuGE (Jones et al., 2007), GelML and mzML have been released or published and, to date, mzData standards for mass spectrometry are widely supported by product manufacturers.

4.1.3 Heterogeneous data integration: Database interoperability

Although standard exchange formats allow easy data exchange between and with repositories, systems level investigations relying on computational analyses require data integration from heterogeneous sources. Instead of trying to duplicate such large amounts of data, integration can be most effectively achieved through connection to repositories. For instance, PeptideAtlas proposes a Distributed Annotation System (DAS) server which allows visualizing data as tracks in the Ensembl Genome Browser (Flicek et al., 2010). Similarly, the PRIDE repository is available through a BioMart service (Smedley et al., 2009). Thus, it can be accessed as a simple REST (Representational State Transfer) web service that involves building an HTTP request including an XML file that encodes the filters and attributes of the request. Web services are indeed being used more and more in

bioinformatics, providing remote access to data and tools that can be combined into workflows with workbench environments like Taverna (Hull et al., 2006) or into client applications (Mcwilliam et al., 2009).

4.2 Data repositories

Proteomics data repositories have been made available to the scientific community on the web since the early nineties. Swiss 2D-PAGE (Hoogland et al., 2004) which collects protein identification from 2DE and 1D-PAGE gels was created in 1993. This database is now part of the World 2D-PAGE portal (Hoogland et al., 2008) which federates 9 gel-based proteomics databases for a total of nearly 18,800 identified spots in 141 maps for 22 species, making it the biggest gel-based proteomics dataset accessible from a single interface (June 2011). Other less general two-dimensional electrophoresis databases are also available and give access to gel-based protein identification in different systems. These repositories provide information on identification data (pI, mW, peptides and spot), links between maps and, of course, link to the identified entries in protein databases. Gels are displayed as an interactive image which can be clicked on to visualize spot information. In addition to electrophoresis data, Proteome Database for Microbial Research (Pleißner et al., 2004) also offers access to MS data.

Proteomics repositories which focus on MS data include GPMDB (Craig et al., 2004), PeptitdeAtlas (Deutsch et al., 2008), PRIDE (Vizcaíno et al., 2009), and Proteome Commons (https://proteomecommons.org/) among the most prominent ones. GPMDB is a relational database that was designed to aid in the process of validating peptide-to-mass spectrum assignment and/or protein coverage patterns. Together with data analysis servers it constitutes the open-source system referred to as the Global Proteome Machine. PeptideAtlas is a multi-organism compendium of raw data coming from high-throughput proteomics technologies. Only raw data are accepted and are periodically reprocessed as more advanced interpretation tools for identification and statistical validation are available. The PeptideAtlas project long-term goal is the annotation of eukaryotic genomes through a robust validation of expressed proteins. PRIDE (Proteomics Identifications Database) is a repository of MS derived peptide and protein identifications, MS mass spectra, and associated metadata. Proteome Commons is a public resource for collaborative research and public sharing of proteomics data, tools and news. Permanent storage of data suitable for publication is provided through a distributed repository. Registered users may set up their own or join group projects for easy collaboration with colleagues or partners as project permissions and member responsibilities can be fine-tuned in order to control access to data. As post-translational modifications like phosphorylation play an important role in control of protein activity some proteomics databases focus on phosphorylation and other PTMs (Phospho.ELM (Dinkel et al., 2011), Phospho3D (Zanzoni et al., 2011), Phosida (Gnad et al., 2011) PhosphoSitePlus® (Hornbeck et al., 2004), and PhosphoPep. (Bodenmiller et al., 2008).

4.3 The need for an environmental proteomics database

As metaproteomics allows to study microorganisms' functionality in their natural context, in the coming years it is likely to become the technique of choice for functional characterization of microbial communities. Moreover, this methodology will give invaluable insights into microbial ecology, in particular when associated with metagenomics data. To correlate the

observed variations in functions with differences in conditions, comparative studies must also consider environment characteristics such as chemical composition (presence of toxic compounds, organic matter), physical properties (temperature), habitat, sample location and collection dates. High-throughput data-mining would therefore require that databases handling environmental metaproteomics include metadata providing environmental information in a computer-readable form. In the absence of a currently available specific standard for environmental proteomics, the content of these metadata could be inspired by corresponding sections of the minimum information about any sequence standard MIxS

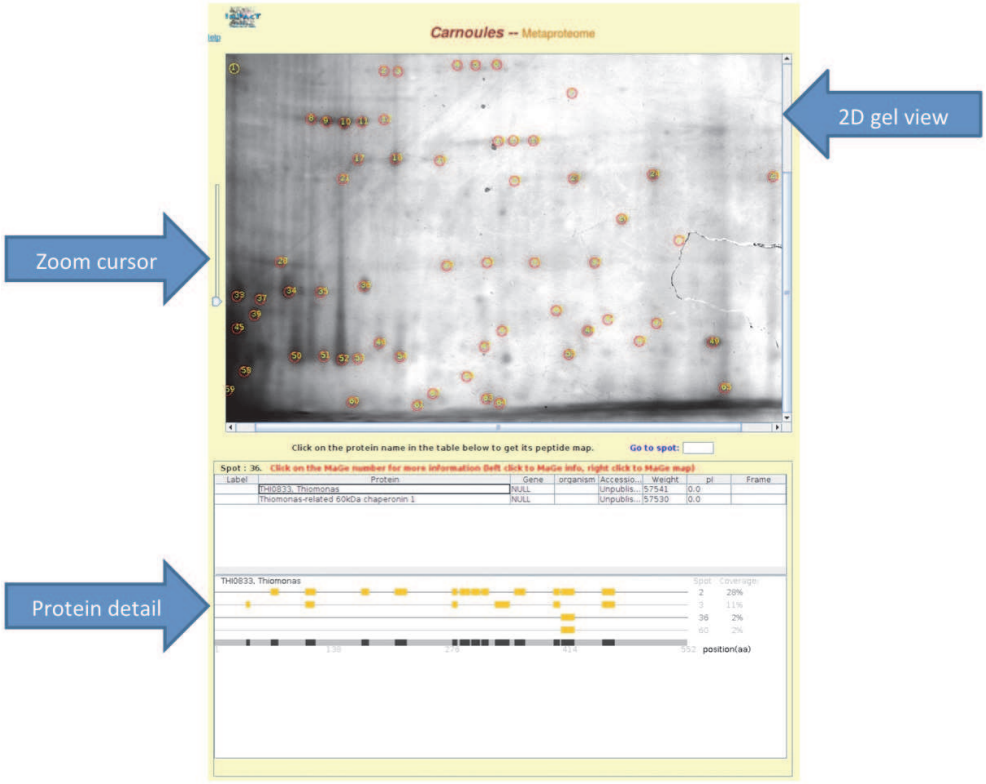


Fig. 3. The InPact proteomic database (http://inpact.u-strasbg.fr). The various functionalities of the interface allow the exploration of specific areas of the 2D gel by using a zoom-in/out function. Spots present in the selected area can be outlined, and the corresponding MS results can be seen for each. In addition, more information can be seen by hovering the mouse over any spot and/or clicking on it (name, Mw, pI, MS peptidic sequence). As an example, one of the numerous GroEL chaperonins identified in the Carnoulès community metaproteome illustrates the data that can be obtained for any protein identified by mass spectrometry, e.g. the label of the corresponding CDS in the genome when available and the spot numbers where the protein has been identified (top), the size and the location within the genome of the MS peptidic fragments obtained as well as their % coverage with respect to the full length CDS.

(Yilmaz et al., 2011) or the IUPAC minimum requirements for reporting analytical data for environmental samples (Egli et al., 2003). For instance, the InPACT environmental microbiology database (http://inpact.u-strasbg.fr/) is a proteomics database dedicated to environmental microbiology providing gel-based data pertaining to microorganisms as well as complex communities (Figure 3). It also provides genomics information thanks to tight links with the MaGe database (Vallenet et al., 2006) and tools for functional profile comparison between gels. InPACT, although still in its infancy, provides tools for gel comparisons at the function level thanks to functional description of proteins using the Gene Ontology (The Gene Ontology Consortium. 2000). Future developments will now focus on the integration of environmental information as metadata and the addition of more comparison tools including multivariate statistical analysis of proteomics data and associated metadata. Integration with external data sources (metabolism, genomic data) will also be reinforced. Finally, in order to increase the accessibility of data, InPACT data access will be offered not only through the web server but also as RESTful web services. In the near future, InPACT and other databases will hopefully prove to be useful proteomics-oriented tools for environmental microbiology.

5. Conclusion

The past few years has seen a huge amount of genomic information published in databases. Associated with functional genomic approaches such as proteomics, those data will greatly improve our knowledge of the structure, the functioning, the diversity and the evolution of microorganisms. Similarly, the study of microbial communities as a whole will be of great interest to investigate complex consortia and to address important questions regarding the role of uncultured microorganisms in microbial ecosystems. Proteomics, when combined not only with other genomic methods such as transcriptomics and metabolomics, but also with more classical methods of genetics, molecular biology and/or biochemistry, will give an integrated view of biological objects present in any environment, their role and their relationships. They will lead to a better understanding of how microorganisms colonize new ecological niches and to the possible use of their specific properties in biotechnology.

6. Acknowledgment

Financial support came from the Université de Strasbourg (UdS), the Centre National de la Recherche Scientifique (EC2CO project and GDR2909 research network, http://gdr2909.alsace.cnrs.fr/) and the Agence Nationale de la Recherche (ANR RARE and MULTIPOLSITE projects).

7. References

Addona, T. A., Abbatiello, S. E., Schilling, B., Skates, S. J., Mani, D. R., Bunk, D. M., Spiegelman, C. H., Zimmerman, L. J., Ham, A.-J. L., Keshishian, H., Hall, S. C., Allen, S., Blackman, R. K., Borchers, C. H., Buck, C., Cardasis, H. L., Cusack, M. P., Dodder, N. G., Gibson, B. W., Held, J. M., Hiltke, T., Jackson, A., Johansen, E. B., Kinsinger, C. R., Li, J., Mesri, M., Neubert, T. A., Niles, R. K., Pulsipher, T. C.,

Ransohoff, D., Rodriguez, H., Rudnick, P. A., Smith, D., Tabb, D. L., Tegeler, T. J., Variyath, A. M., Vega-Montoto, L. J., Wahlander, A., Waldemarson, S., Wang, M., Whiteaker, J. R., Zhao, L., Anderson, N. L., Fisher, S. J., Liebler, D. C., Paulovich, A. G., Regnier, F. E., Tempst, P., & Carr, S. A. (2009). Multi-site assessment of the precision and reproducibility of multiple reaction monitoring-based measurements of proteins in plasma. *Nature Biotechnology*, Vol. 27, No. 7, pp. 633-641.

Appel, R. D., Bairoch, A., Sanchez, J. C., Vargas, J. R., Golaz, O., Pasquali, C., & Hochstrasser, D. F. (1996). Federated two-dimensional electrophoresis database: a simple means of publishing two-dimensional electrophoresis data. *Electrophoresis*, Vol. 17, No. 3, pp. 540-546.

Belnap, C. P., Pan, C., Denef, V. J., Samatova, N. F., Hettich, R. L., & Banfield, J. F. (2011). Quantitative proteomic analyses of the response of acidophilic microbial communities to different pH conditions. *The ISME Journal*, Vol. 5, No. 7, pp. 1152-1161.

Benndorf, D., Balcke, G. U., Harms, H., & von Bergen, M. (2007). Functional metaproteome analysis of protein extracts from contaminated soil and groundwater. *The ISME Journal*, Vol. 1, No. 3, pp. 224-234.

Bertin, P. N., Heinrich-Salmeron, A., Pelletier, E., Goulhen-Chollet, F., Arsène-Ploetze, F., Gallien, S., Lauga, B., Casiot, C., Calteau, A., Vallenet, D., Bonnefoy, V., Bruneel, O., Chane-Woon-Ming, B., Cleiss-Arnold, J., Duran, R., Elbaz-Poulichet, F., Fonknechten, N., Giloteaux, L., Halter, D., Koechler, S., Marchal, M., Mornico, D., Schaeffer, C., Smith, A. A. T., Van Dorsselaer, A., Weissenbach, J., Médigue, C., & Le Paslier, D. (2011). Metabolic diversity among main microorganisms inside an arsenic-rich ecosystem revealed by meta- and proteo-genomics. *International Society for Microbial Ecology*, In press.

Bertin, P. N., Medigue, C., & Normand, P. (2008). Advances in environmental genomics: towards an integrated view of microorganisms and ecosystems. *Microbiology*, Vol. 154, No. 2, pp. 347-359.

Beynon, R. J., Doherty, M. K., Pratt, J. M., & Gaskell, S. J. (2005). Multiplexed absolute quantification in proteomics using artificial QCAT proteins of concatenated signature peptides. *Nature Methods*, Vol. 2, No. 8, pp. 587-589.

Bodenmiller, B., Campbell, D., Gerrits, B., Lam, H., Jovanovic, M., Picotti, P., Schlapbach, R., & Aebersold, R. (2008). PhosphoPep--a database of protein phosphorylation sites in model organisms. *Nature Biotechnology*, Vol. 26, No. 12, pp. 1339-1340.

Bona, E., Marsano, F., Massa, N., Cattaneo, C., Cesaro, P., Argese, E., di Toppi, L. S., Cavaletto, M., & Berta, G. (2011). Proteomic analysis as a tool for investigating arsenic stress in *Pteris vittata* roots colonized or not by arbuscular mycorrhizal symbiosis. *Journal of Proteomics*. In press.

Brun, V., Dupuis, A., Adrait, A., Marcellin, M., Thomas, D., Court, M., Vandenesch, F., & Garin, J. (2007). Isotope-labeled protein standards: toward absolute quantitative proteomics. *Molecular & Cellular Proteomics: MCP*, Vol. 6, No. 12, pp. 2139-2149.

Bruneel, O., Volant, A., Gallien, S., Chaumande, B., Casiot, C., Carapito, C., Bardil, A., Morin, G., Brown, G. E., Personné, C. J., Le Paslier, D., Schaeffer, C., Van Dorsselaer, A., Bertin, P. N., Elbaz-Poulichet, F., & Arsène-Ploetze, F. (2011). Characterization of the Active Bacterial Community Involved in Natural

Attenuation Processes in Arsenic-Rich Creek Sediments. *Microbial Ecology*, Vol. 61, No. Issue 4, pp. 793-810.

Bryan, C. G., Marchal, M., Battaglia-Brunet, F., Kugler, V., Lemaitre-Guillier, C., Lièvremont, D., Bertin, P. N., & Arsène-Ploetze, F. (2009). Carbon and arsenic metabolism in *Thiomonas* strains: differences revealed diverse adaptation processes. *BMC Microbiology*, Vol. 9, pp. 127.

Callister, S. J., Wilkins, M. J., Nicora, C. D., Williams, K. H., Banfield, J. F., VerBerkmoes, N. C., Hettich, R. L., N'Guessan, L., Mouser, P. J., Elifantz, H., Smith, R. D., Lovley, D. R., Lipton, M. S., & Long, P. E. (2010). Analysis of biostimulated microbial communities from two field experiments reveals temporal and spatial differences in proteome profiles. *Environmental Science & Technology*, Vol. 44, No. 23, pp. 8897-8903.

Cannon, W., & Webb-Robertson, B.-J. (2007). Computational proteomics : High-throughput analysis for Systems Biology, *Proceedings of Pacific Symposium on Biocomputing*, Vol. 12, p. 403-408.

Carapito, C., Muller, D., Turlin, E., Koechler, S., Danchin, A., Van Dorsselaer, A., Leize-Wagner, E., Bertin, P. N., & Lett, M.-C. (2006). Identification of genes and proteins involved in the pleiotropic response to arsenic stress in *Caenibacter arsenoxydans*, a metalloresistant beta-proteobacterium with an unsequenced genome. *Biochimie*, Vol. 88, No. 6, pp. 595-606.

Casado-Vela, J., Cebrián, A., del Pulgar, M. T. G., Sánchez-López, E., Vilaseca, M., Menchén, L., Diema, C., Sellés-Marchart, S., Martínez-Esteso, M. J., Yubero, N., Bru-Martínez, R., Lacal, J. C., & Lacal, J. C. (2011). Lights and shadows of proteomic technologies for the study of protein species including isoforms, splicing variants and protein post-translational modifications. *Proteomics*, Vol. 11, No. 4, pp. 590-603.

Cañas, B., Piñeiro, C., Calvo, E., López-Ferrer, D., & Gallardo, J. M. (2007). Trends in sample preparation for classical and second generation proteomics. *Journal of Chromatography. A*, Vol. 1153, No. 1-2, pp. 235-258.

Cleiss-Arnold, J., Koechler, S., Proux, C., Fardeau, M.-L., Dillies, M.-A., Coppee, J.-Y., Arsène-Ploetze, F., & Bertin, P. N. (2010). Temporal transcriptomic response during arsenic stress in *Herminiimonas arsenicoxydans*. *BMC Genomics*, Vol. 11, pp. 709.

Craig, R., Cortens, J. P., & Beavis, R. C. (2004). Open Source System for Analyzing, Validating, and Storing Protein Identification Data. *Journal of Proteome Research*, Vol. 3, No. 6, pp. 1234-1242.

Delalande, F., Carapito, C., Brizard, J.-P., Brugidou, C., & Van Dorsselaer, A. (2005). Multigenic families and proteomics: extended protein characterization as a tool for paralog gene identification. *Proteomics*, Vol. 5, No. 2, pp. 450-460.

Denef, V. J., Kalnejais, L. H., Mueller, R. S., Wilmes, P., Baker, B. J., Thomas, B. C., VerBerkmoes, N. C., Hettich, R. L., & Banfield, J. F. (2010a). Proteogenomic basis for ecological divergence of closely related bacteria in natural acidophilic microbial communities. *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 107, No. 6, pp. 2383-2390.

Denef, V. J., Mueller, R. S., & Banfield, J. F. (2010b). AMD biofilms: using model communities to study microbial evolution and ecological complexity in nature. *The ISME Journal*, Vol. 4, No. 5, pp. 599-610.

Deutsch, E. W., Lam, H., & Aebersold, R. (2008). PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO reports*, Vol. 9, No. 5, pp. 429-434.

Dinkel, H., Chica, C., Via, A., Gould, C. M., Jensen, L. J., Gibson, T. J., & Diella, F. (2011). Phospho.ELM: a database of phosphorylation sites--update 2011. *Nucleic Acids Research*, Vol. 39, No. Database issue, pp. D261-267.

Egli, H., Dassenakis, M., Garelick, H., Van Grieken, R., Peijnenburg, W. J. G. M., Klasinc, L., Kördel, W., Priest, N., & Tavares, T. (2003). Minimum requirements for reporting analytical data for environmental samples (IUPAC Technical Report). *Pure and Applied Chemistry*, Vol. 75, No. 8, pp. 1097-1106.

Elschenbroich, S., & Kislinger, T. (2011). Targeted proteomics by selected reaction monitoring mass spectrometry: applications to systems biology and biomarker discovery. *Molecular bioSystems*, Vol. 7, No. 2, pp. 292-303.

Falkner, J. A., Hill, J. A., & Andrews, P. C. (2008). Proteomics FASTA Archive and Reference Resource. *Proteomics*, Vol. 8, No. 9, pp. 1756-1757.

Ferrer, M., Golyshina, O. V., Beloqui, A., Golyshin, P. N., & Timmis, K. N. (2007). The cellular machinery of *Ferroplasma acidiphilum* is iron-protein-dominated. *Nature*, Vol. 445, No. 7123, pp. 91-94.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., et al. (2010). Ensembl 2011. *Nucleic Acids Research*, Vol. 39, No. Database, pp. D800-D806.

Fränzel, B., & Wolters, D. A. (2011). Advanced MudPIT as a next step towards high proteome coverage. *Proteomics*. In press.

Gallien, S., Perrodou, E., Carapito, C., Deshayes, C., Reyrat, J.-M., Van Dorsselaer, A., Poch, O., Schaeffer, C., & Lecompte, O. (2009). Ortho-proteogenomics: multiple proteomes investigation through orthology and a new MS-based protocol. *Genome Research*, Vol. 19, No. 1, pp. 128-135.

Gevaert, K., Impens, F., Ghesquière, B., Van Damme, P., Lambrechts, A., & Vandekerckhove, J. (2008). Stable isotopic labeling in proteomics. *Proteomics*, Vol. 8, No. 23-24, pp. 4873-4885.

Gnad, F., Gunawardena, J., & Mann, M. (2011). PHOSIDA 2011: the posttranslational modification database. *Nucleic Acids Research*, Vol. 39, No. Database issue, pp. D253-260.

Grangeasse, C., Terreux, R., & Nessler, S. (2010). Bacterial tyrosine-kinases: structure-function analysis and therapeutic potential. *Biochimica Et Biophysica Acta*, Vol. 1804, No. 3, pp. 628-634.

Gundry, R. L., White, M. Y., Murray, C. I., Kane, L. A., Fu, Q., Stanley, B. A., & Van Eyk, J. E. (2009). Preparation of proteins and peptides for mass spectrometry analysis in a bottom-up proteomics workflow. *Current Protocols in Molecular Biology,* Frederick M. Ausubel*,* Chapter 10, pp. Unit 10.25.

Halter, D., Cordi, A., Gribaldo, S., Gallien, S., Goulhen-Chollet, F., Heinrich-Salmeron, A., Carapito, C., Pagnout, C., Montaut, D., Seby, F., Van Dorsselaer, A., Schaeffer, C., Bertin, P. N., Bauda, P., & Arsène-Ploetze, F. (2011). Taxonomic and functional

prokaryote diversity in mildly arsenic-contaminated sediments. *Research in Microbiology*. In press.

Hecker, M., & Völker, U. (2004). Towards a comprehensive understanding of *Bacillus subtilis* cell physiology by physiological proteomics. *Proteomics*, Vol. 4, No. 12, pp. 3727-3750.

Hoogland, C., Mostaguir, K., Appel, R., & Lisacek, F. (2008). The World-2DPAGE Constellation to promote and publish gel-based proteomics data through the ExPASy server. *Journal of Proteomics*, Vol. 71, No. 2, pp. 245-248.

Hoogland, C., Mostaguir, K., Sanchez, J.-C., Hochstrasser, D. F., & Appel, R. D. (2004). SWISS-2DPAGE, ten years later. *Proteomics*, Vol. 4, No. 8, pp. 2352-2356.

Hornbeck, P. V., Chabra, I., Kornhauser, J. M., Skrzypek, E., & Zhang, B. (2004). PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation. *Proteomics*, Vol. 4, No. 6, pp. 1551-1561.

Hull, D., Wolstencroft, K., Stevens, R., Goble, C., Pocock, M. R., Li, P., & Oinn, T. (2006). Taverna: a tool for building and running workflows of services. *Nucleic Acids Research*, Vol. 34, No. Web Server, pp. W729-W732.

Hüttenhain, R., Malmström, J., Picotti, P., & Aebersold, R. (2009). Perspectives of targeted mass spectrometry for protein biomarker verification. *Current Opinion in Chemical Biology*, Vol. 13, No. 5-6, pp. 518-525.

Jones, A. R., Miller, M., Aebersold, R., Apweiler, R., Ball, C. A., Brazma, A., DeGreef, J., Hardy, N., Hermjakob, H., Hubbard, S. J., Hussey, P., Igra, M., Jenkins, H., Julian, R. K., Laursen, K., Oliver, S. G., Paton, N. W., Sansone, S.-A., Sarkans, U., Stoeckert, C. J., Taylor, C. F., Whetzel, P. L., White, J. A., Spellman, P., & Pizarro, A. (2007). The Functional Genomics Experiment model (FuGE): an extensible framework for standards in functional genomics. *Nature Biotechnology*, Vol. 25, No. 10, pp. 1127-1133.

Jungblut, P. R. (2001). Proteome analysis of bacterial pathogens. *Microbes and Infection / Institut Pasteur*, Vol. 3, No. 10, pp. 831-840.

Kan, J., Hanson, T. E., Ginter, J. M., Wang, K., & Chen, F. (2005). Metaproteomic analysis of Chesapeake Bay microbial communities. *Saline Systems*, Vol. 1, pp. 7.

Keshishian, H., Addona, T., Burgess, M., Mani, D. R., Shi, X., Kuhn, E., Sabatine, M. S., Gerszten, R. E., & Carr, S. A. (2009). Quantification of cardiovascular biomarkers in patient plasma by targeted mass spectrometry and stable isotope dilution. *Molecular & Cellular Proteomics: MCP*, Vol. 8, No. 10, pp. 2339-2349.

Kim, Y. H., Cho, K., Yun, S.-H., Kim, J. Y., Kwon, K.-H., Yoo, J. S., & Kim, S. I. (2006). Analysis of aromatic catabolic pathways in *Pseudomonas putida* KT 2440 using a combined proteomic approach: 2-DE/MS and cleavable isotope-coded affinity tag analysis. *Proteomics*, Vol. 6, No. 4, pp. 1301-1318.

Lacerda, C. M. R., Choe, L. H., & Reardon, K. F. (2007). Metaproteomic analysis of a bacterial community response to cadmium exposure. *Journal of Proteome Research*, Vol. 6, No. 3, pp. 1145-1152.

Laemmli, U. K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature*, Vol. 227, No. 5259, pp. 680-685.

Lange, V., Picotti, P., Domon, B., & Aebersold, R. (2008). Selected reaction monitoring for quantitative proteomics: a tutorial. *Molecular Systems Biology*, Vol. 4, pp. 222.

Liedert, C., Bernhardt, J., Albrecht, D., Voigt, B., Hecker, M., Salkinoja-Salonen, M., & Neubauer, P. (2010). Two-dimensional proteome reference map for the radiation-resistant bacterium *Deinococcus geothermalis*. *Proteomics*, Vol. 10, No. 3, pp. 555-563.

Linares, J. F., Moreno, R., Fajardo, A., Martínez-Solano, L., Escalante, R., Rojo, F., & Martínez, J. L. (2010). The global regulator Crc modulates metabolism, susceptibility to antibiotics and virulence in *Pseudomonas aeruginosa*. *Environmental Microbiology*, Vol. 12, No. 12, pp. 3196-3212.

Mary, I., Oliver, A., Skipp, P., Holland, R., Topping, J., Tarran, G., Scanlan, D. J., O'Connor, C. D., Whiteley, A. S., Burkill, P. H., & Zubkov, M. V. (2010). Metaproteomic and metagenomic analyses of defined oceanic microbial populations using microwave cell fixation and flow cytometric sorting. *FEMS Microbiology Ecology*, Vol. 74, No. 1, pp. 10-18.

Mcwilliam, H., Valentin, F., Goujon, M., Li, W., Narayanasamy, M., Martin, J., Miyar, T., & Lopez, R. (2009). Web services at the European Bioinformatics Institute-2009. *Nucleic Acids Research*, Vol. 37, No. Web Server, pp. W6-W10.

Monteiro, K. M., de Carvalho, M. O., Zaha, A., & Ferreira, H. B. (2010). Proteomic analysis of the *Echinococcus granulosus* metacestode during infection of its intermediate host. *Proteomics*, Vol. 10, No. 10, pp. 1985-1999.

Moretti, M., Grunau, A., Minerdi, D., Gehrig, P., Roschitzki, B., Eberl, L., Garibaldi, A., Gullino, M. L., & Riedel, K. (2010). A proteomics approach to study synergistic and antagonistic interactions of the fungal-bacterial consortium *Fusarium oxysporum* wild-type MSA 35. *Proteomics*, Vol. 10, No. 18, pp. 3292-3320.

Morris, R. M., Nunn, B. L., Frazar, C., Goodlett, D. R., Ting, Y. S., & Rocap, G. (2010). Comparative metaproteomics reveals ocean-scale shifts in microbial nutrient utilization and energy transduction. *The ISME Journal*, Vol. 4, No. 5, pp. 673-685.

Mostaguir, K., Hoogland, C., Binz, P.-A., & Appel, R. D. (2003). The Make 2D-DB II package: conversion of federated two-dimensional gel electrophoresis databases into a relational format and interconnection of distributed databases. *Proteomics*, Vol. 3, No. 8, pp. 1441-1444.

Mueller, R. S., Denef, V. J., Kalnejais, L. H., Suttle, K. B., Thomas, B. C., Wilmes, P., Smith, R. L., Nordstrom, D. K., McCleskey, R. B., Shah, M. B., Verberkmoes, N. C., Hettich, R. L., & Banfield, J. F. (2010). Ecological distribution and population physiology defined by proteomics in a natural microbial community. *Molecular Systems Biology*, Vol. 6, pp. 374.

Muller, D., Médigue, C., Koechler, S., Barbe, V., Barakat, M., Talla, E., Bonnefoy, V., Krin, E., Arsène-Ploetze, F., Carapito, C., Chandler, M., Cournoyer, B., Cruveiller, S., Dossat, C., Duval, S., Heymann, M., Leize, E., Lieutaud, A., Lièvremont, D., Makita, Y., Mangenot, S., Nitschke, W., Ortet, P., Perdrial, N., Schoepp, B., Siguier, P., Simeonova, D. D., Rouy, Z., Segurens, B., Turlin, E., Vallenet, D., Van Dorsselaer, A., Weiss, S., Weissenbach, J., Lett, M.-C., Danchin, A., & Bertin, P. N. (2007). A tale of two oxidation states: bacterial colonization of arsenic-rich environments. *PLoS Genetics*, Vol. 3, No. 4, pp. e53.

Natale, D. A., Arighi, C. N., Barker, W. C., Blake, J. A., Bult, C. J., Caudy, M., Drabkin, H. J., D'Eustachio, P., Evsikov, A. V., Huang, H., Nchoutmboube, J., Roberts, N. V., Smith, B., Zhang, J., & Wu, C. H. (2010). The Protein Ontology: a structured

representation of protein forms and complexes. *Nucleic Acids Research*, Vol. 39, No. Database, pp. D539-D545.

Nesvizhskii, A. I. (2010). A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *Journal of Proteomics*, Vol. 73, No. 11, pp. 2092-2123.

Orchard, S., & Hermjakob, H. (2007). The HUPO proteomics standards initiative--easing communication and minimizing data loss in a changing world. *Briefings in Bioinformatics*, Vol. 9, No. 2, pp. 166-173.

Picotti, P., Bodenmiller, B., Mueller, L. N., Domon, B., & Aebersold, R. (2009). Full dynamic range proteome analysis of *S. cerevisiae* by targeted proteomics. *Cell*, Vol. 138, No. 4, pp. 795-806.

Picotti, P., Rinner, O., Stallmach, R., Dautel, F., Farrah, T., Domon, B., Wenschuh, H., & Aebersold, R. (2010). High-throughput generation of selected reaction-monitoring assays for proteins and proteomes. *Nature Methods*, Vol. 7, No. 1, pp. 43-46.

Piette, F., D'Amico, S., Struvay, C., Mazzucchelli, G., Renaut, J., Tutino, M. L., Danchin, A., Leprince, P., & Feller, G. (2010). Proteomics of life at low temperatures: trigger factor is the primary chaperone in the Antarctic bacterium *Pseudoalteromonas haloplanktis* TAC125. *Molecular Microbiology*, Vol. 76, No. 1, pp. 120-132.

Pleißner, K.-P., Eifert, T., Buettner, S., Schmidt, F., Boehme, M., Meyer, T. F., Kaufmann, S. H. E., & Jungblut, P. R. (2004). Web-accessible proteome databases for microbial research. *Proteomics*, Vol. 4, No. 5, pp. 1305-1313.

Rabilloud, T., Chevallet, M., Luche, S., & Lelong, C. (2010). Two-dimensional gel electrophoresis in proteomics: Past, present and future. *Journal of Proteomics*, Vol. 73, No. 11, pp. 2064-2077.

Ram, R. J., Verberkmoes, N. C., Thelen, M. P., Tyson, G. W., Baker, B. J., Blake, R. C., 2nd, Shah, M., Hettich, R. L., & Banfield, J. F. (2005). Community proteomics of a natural microbial biofilm. *Science (New York, N.Y.)*, Vol. 308, No. 5730, pp. 1915-1920.

Ramabu, S. S., Ueti, M. W., Brayton, K. A., Baszler, T. V., & Palmer, G. H. (2010). Identification of *Anaplasma marginale* proteins specifically upregulated during colonization of the tick vector. *Infection and Immunity*, Vol. 78, No. 7, pp. 3047-3052.

Saunders, N. F. W., Goodchild, A., Raftery, M., Guilhaus, M., Curmi, P. M. G., & Cavicchioli, R. (2005). Predicted roles for hypothetical proteins in the low-temperature expressed proteome of the Antarctic archaeon *Methanococcoides burtonii*. *Journal of Proteome Research*, Vol. 4, No. 2, pp. 464-472.

Schleifer, K. H. (2009). Classification of Bacteria and Archaea: past, present and future. *Systematic and Applied Microbiology*, Vol. 32, No. 8, pp. 533-542.

Schmidt, F., Donahoe, S., Hagens, K., Mattow, J., Schaible, U. E., Kaufmann, S. H. E., Aebersold, R., & Jungblut, P. R. (2004). Complementary Analysis of the *Mycobacterium tuberculosis* Proteome by Two-dimensional Electrophoresis and Isotope-coded Affinity Tag Technology. *Molecular & Cellular Proteomics*, Vol. 3, No. 1, pp. 24 -42.

Sengupta, N., & Alam, S. I. (2011). In vivo studies of *Clostridium perfringens* in mouse gas gangrene model. *Current Microbiology*, Vol. 62, No. 3, pp. 999-1008.

Simmons, S. L., Dibartolo, G., Denef, V. J., Goltsman, D. S. A., Thelen, M. P., & Banfield, J. F. (2008). Population genomic analysis of strain variation in *Leptospirillum* group II

bacteria involved in acid mine drainage formation. *PLoS Biology*, Vol. 6, No. 7, pp. e177.

Smedley, D., Haider, S., Ballester, B., Holland, R., London, D., Thorisson, G., & Kasprzyk, A. (2009). BioMart--biological queries made easy. *BMC Genomics*, Vol. 10, pp. 22.

Taylor, C. F., Paton, N. W., Lilley, K. S., Binz, P.-A., Julian, R. K., Jones, A. R., Zhu, W., Apweiler, R., Aebersold, R., Deutsch, E. W., Dunn, M. J., Heck, A. J. R., Leitner, A., Macht, M., Mann, M., Martens, L., Neubert, T. A., Patterson, S. D., Ping, P., Seymour, S. L., Souda, P., Tsugita, A., Vandekerckhove, J., Vondriska, T. M., Whitelegge, J. P., Wilkins, M. R., Xenarios, I., Yates, J. R., & Hermjakob, H. (2007). The minimum information about a proteomics experiment (MIAPE). *Nature Biotechnology*, Vol. 25, No. 8, pp. 887-893.

Taylor, E. B., & Williams, M. A. (2010). Microbial protein in soil: influence of extraction method and C amendment on extraction and recovery. *Microbial Ecology*, Vol. 59, No. 2, pp. 390-399.

Vallenet, D., Labarre, L., Rouy, Z., Barbe, V., Bocs, S., Cruveiller, S., Lajus, A., Pascal, G., Scarpelli, C., & Médigue, C. (2006). MaGe: a microbial genome annotation system supported by synteny results. *Nucleic Acids Research*, Vol. 34, No. 1, pp. 53-65.

Vizcaíno, J. A., Côté, R., Reisinger, F., M. Foster, J., Mueller, M., Rameseder, J., Hermjakob, H., & Martens, L. (2009). A guide to the Proteomics Identifications Database proteomics data repository. *Proteomics*, Vol. 9, No. 18, pp. 4276-4283.

Weiss, S., Carapito, C., Cleiss, J., Koechler, S., Turlin, E., Coppee, J.-Y., Heymann, M., Kugler, V., Stauffert, M., Cruveiller, S., Médigue, C., Van Dorsselaer, A., Bertin, P. N., & Arsène-Ploetze, F. (2009). Enhanced structural and functional genome elucidation of the arsenite-oxidizing strain *Herminiimonas arsenicoxydans* by proteomics data. *Biochimie*, Vol. 91, No. 2, pp. 192-203.

Werner, J. J., Ptak, A. C., Rahm, B. G., Zhang, S., & Richardson, R. E. (2009). Absolute quantification of *Dehalococcoides* proteins: enzyme bioindicators of chlorinated ethene dehalorespiration. *Environmental Microbiology*, Vol. 11, No. 10, pp. 2687-2697.

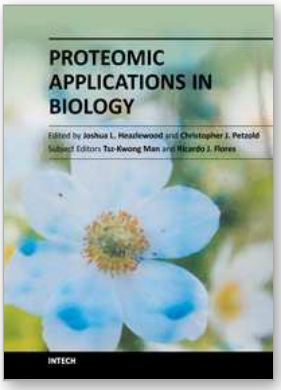
Wilkins, M. J., Verberkmoes, N. C., Williams, K. H., Callister, S. J., Mouser, P. J., Elifantz, H., N'guessan, A. L., Thomas, B. C., Nicora, C. D., Shah, M. B., Abraham, P., Lipton, M. S., Lovley, D. R., Hettich, R. L., Long, P. E., & Banfield, J. F. (2009). Proteogenomic monitoring of *Geobacter* physiology during stimulated uranium bioremediation. *Applied and Environmental Microbiology*, Vol. 75, No. 20, pp. 6591-6599.

Wilmes, P., & Bond, P. L. (2004). The application of two-dimensional polyacrylamide gel electrophoresis and downstream analyses to a mixed community of prokaryotic microorganisms. *Environmental Microbiology*, Vol. 6, No. 9, pp. 911-920.

Yan, J. X., Devenish, A. T., Wait, R., Stone, T., Lewis, S., & Fowler, S. (2002). Fluorescence two-dimensional difference gel electrophoresis and mass spectrometry based proteomic analysis of *Escherichia coli*. *Proteomics*, Vol. 2, No. 12, pp. 1682-1698.

Yilmaz, P., Kottmann, R., Field, D., Knight, R., Cole, J. R., et al. (2011). Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nature Biotechnology*, Vol. 29, No. 5, pp. 415-420.

Zanzoni, A., Carbajo, D., Diella, F., Gherardini, P. F., Tramontano, A., Helmer-Citterich, M., & Via, A. (2011). Phospho3D 2.0: an enhanced database of three-dimensional structures of phosphorylation sites. *Nucleic Acids Research*, Vol. 39, No. Database issue, pp. D268-271.



Proteomic Applications in Biology

Edited by Dr. Joshua Heazlewood

ISBN 978-953-307-613-3
Hard cover, 264 pages
Publisher InTech
Published online 18, January, 2012
Published in print edition January, 2012

The past decade has seen the field of proteomics expand from a highly technical endeavor to a widely utilized technique. The objective of this book is to highlight the ways in which proteomics is currently being employed to address issues in the biological sciences. Although there have been significant advances in techniques involving the utilization of proteomics in biology, fundamental approaches involving basic sample visualization and protein identification still represent the principle techniques used by the vast majority of researchers to solve problems in biology. The work presented in this book extends from overviews of proteomics in specific biological subject areas to novel studies that have employed a proteomics-based approach. Collectively they demonstrate the power of established and developing proteomic techniques to characterize complex biological systems.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Florence Arsène-Ploetze, Christine Carapito, Frédéric Plewniak and Philippe N. Bertin (2012). Proteomics as a Tool for the Characterization of Microbial Isolates and Complex Communities, Proteomic Applications in Biology, Dr. Joshua Heazlewood (Ed.), ISBN: 978-953-307-613-3, InTech, Available from: http://www.intechopen.com/books/proteomic-applications-in-biology/proteomics-as-a-tool-for-the-characterization-of-microbial-isolates-and-complex-communities

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821